Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 08, August-2020

Comparison of Matrix Factorization and Graph Based Models for Summay Extraction

Harneet Singh
Department of Information Technology
Maharaja Agrasen Institute of Technology
New Delhi, India

***Abstract*—Both Matrix Factorization Based and Graph Based Models are unsupervised models for extracting summary from text documents. As training phase is not required for both these methods , they are extremely fast. Both these methods construct an intermediate representation of text document and use it to assign a score to each sentence present in the text document. In this paper we discuss the underlying concept behind both the methods and compare them on quality of summary extracted.**

Keywords—Text document , matrix factorization , graph model , summary , key-phrase ,latent structure .

I. INTRODUCTION

Text summarization [1] is becoming a popular tool for processing large amount of text information that is available in digital format. Online user participation has increased and more people are contributing towards digital content generation.Text corpus have grown massively and it is very difficult to search for relevant information. The main purpose of using text summarization is to process these text documents and provide relevant results in concise format that can be easily interpreted by humans e.g. keywords,summary. A summary [2] contains only a small subset of sentences of entire document but captures the main theme of document and also reduces the reading time. The user can use this summary to make a decision whether the text document is relevant for his/her work or not.

There are two types of summarization methods , one is extractive and other is abstractive. In extractive based summarization we use existing sentences in text document to create summary. For this task we use mathematical and statistical models to find out which phrases and sentences are important.

Abstractive [3][4] methods use Natural Language Processing techniques to generate a semantic representation of document and use it to extract meaningful constructs from text document and create summary by using new sentences not present in original document. Abstractive methods are far more complex than extractive methods. The models discussed in this paper are extractive methods.

Extractive methods [5] are based on extracting important portions of text using some mathematical or statistical model and using features such as frequency of words,phrases ,location of cue word. Generally these methods use heuristics like more frequently occuring words, phrases or most favourably positioned sentences. Unlike humans these methods don't understand the semantic meaning of text. The advantage of these methods is that they are simple to understand and easy to implement.Extractive methods [6] require some preprocessing before actual algorithm can be applied.

It is very important to understand the importance of text pre-processing. Text is present in highly unstructured format. If we directly apply text mining algorithms on it , we will fail to capture the relevant insights. In any machine learning task , data preprocessing is a crucial step . Not only it standardizes the data , it makes it suitable for machine learning algorithms and also improves the accuracy of these algorithms.

Text data contains noise in form of special characters, delimiters (like comma , full stop , hyphens etc) . These are meaningless for perspective of text summarization. Also various frequently occuring words like a , an ,the , is etc doesn’t tell anything about the topics or major theme of the document. Also a text may contain various forms of same words , text summarization algorithm will consider all of those occurences to be different words but in reality they are same. We need to convert all these forms to a base form.

Sometimes we want to work with certain phrases like noun-verb or noun-verb-adjective , these require special type of preprocessing . First we need to identify what category each word belongs to and then use a pattern to recognize the required phrases.

Pre Processing generally includes but is not limited to: a) Sentences boundary identification. In English, sentence boundary is identified with presence of full stop or semi colon . b) Stop-Word Removal—Frequently occuring words with no semantics and which do not contain relevant information are eliminated. c) Stemming/Lemmatization—The purpose is to obtain the stem or root form of each word, which emphasize its semantics. In Processing step, features influencing the significance of sentences are decided and quantified and then scores are assigned to these features . Top ranked sentences are selected for final summary.

There are some problems [7][8] associated with extractive methods : 1. Main content of document is generally spread across sentences/paragraphs , this cannot be captured if there is restriction on size of summary required. 2. Because we are restricted to use the same sentences present in text document the resulting summary may not be coherent. 3. The sentences chosen for summary are generally longer because part of sentences that are not relevant for summary are also included.4. Different sentences occur in different content , they are just extracted and concatenated in summary , hence they may not be semantically meaningful. These problems are most severe in when summary is extracted from multiple documents/sources.

II. GRAPH BASED MODEL

In graph based models the text document is represented as graph and computations are performed on this graph to assign relevance scores to sentences. To create a graph , generally phrases/sentences containing nouns and verbs are extracted from pre processed text. Each phrase/sentence is added as a node in the graph. The edge between nodes represent some sort of relationship between phrases/sentences like content overlap or similarity. Weight of edge is a measure of degree of relationship , if this weight is below a certain threshold the edge is not created. In graph occurence of dense subgraphs represent different topics.

After creation of graph , some computations are performed to assign scores to each node. The computation is designed in such as manner that nodes having more neighbors are favoured and nodes connected to high scoring nodes are also favoured. Such nodes receive high increment in scores when values are updated. Such computations are performed until scores become stable.

The is also a stochastic perspective to interpret the graph model. Weights of edges can be scaled so that for a given node , the sum of weights of all edges is one. Now this becomes a probability distribution. Graph can be interpreted as Markov Chain and the weights of edge represent probability of transition from one state (in this case a vertex) to another. This opens door for applying techniques of stochastic processes to find out probability of ending up at a particular node at any given time after undergoing a number of transitions from a start state. As number of transitions mode is increased , the probability of being present at each vertex becomes stable.Now nodes with high probability are the most important nodes and sentences corresponding to these nodes will be contributing towards summary.

A big advantage [9][10] of graph based model is that it is suitable for both single document and multi document summarization. Another advantage [11] is that it doesn't need any complex language specific processing and hence can be used for languages other than English.

III. MATRIX FACTORIZATION MODEL

Singular Value Decomposition is a popular matrix factorization method that is used in text mining and is known by Latent Semantic Analysis. It is called latent semantic analysis because of it's ability to capture hidden semantic similarity between sentences even if those sentences don't share any common word but occur in similar context. Another important property of LSA is that it is able to quantify the weightage of each topic in the document in form of a diagonal matrix . This information is useful for determining important topics to be captured in summary and designing a mechanism to favour sentences that relate to these significant topics.

Gong and Liu [12] proposed LSA for single and multi document summarization of news. First a matrix representation of text document is created using and is called term document matrix. This matrix contains Tf-Idf of each word in a sentence . Frequency or binary occurence can also be used instead of tfidf. This matrix is factorized into three matrices generally denoted by U,Σ,V^T.

Matrix U is called left singular matrix and each column of U is interpreted as a topic. The words corresponding to a topic will have high weight in column of that topic. Matrix Σ is called singular matrix and is a diagonal matrix. The diagonal values are called singular values and are sorted. Each singular value represent the weight of topic , jth singular value corresponds to weight of topic denoted by jth column of matrix U. Generally topic having weight below a certain threshold are ignored. matrix V^T is called right singular matrix and contains weight of topics in each sentence. Matrix Σ is multiplied by matrix V^T to get the significance of each sentence.

IV. EXPERIMENT

For matrix based model we have implemented Latent Semantic Analysis and for graph based model text rank is used. Various text samples were collected from online sources and were used for summarization. Text pre-processing included tokenization , stopword removal and lemmatization. For constructing term-document matrix we used three parameters which are binary occurence , word frequency and tfidf.

A. Sample Text 1

In economics, finance is a field that is concerned with the allocation (investment) of assets and liabilities over space and time,often under conditions of risk or uncertainty. Finance can also be defined as the science of money management. Participants in the market aim to price assets based on their risk level, fundamental value, and their expected rate of return. Financial economics is the branch of economics studying the interrelation of financial variables, such as prices, interest rates and shares, as opposed to goods and services. Financial economics concentrates on influences of real economic variables on financial ones, in contrast to pure finance. It centres on managing risk in the context of the financial markets, and the resultant economic and financial models. It essentially explores how rational investors would apply risk and return to the problem of an investment policy. Here, the twin assumptions of rationality and market efficiency lead to modern portfolio theory (the CAPM), and to the Black–Scholes theory for option valuation; it further studies phenomena and models where these assumptions do not hold, or are extended. "Financial economics", at least formally, also considers investment under "certainty" and hence also contributes to corporate finance theory. Financial econometrics is the branch of financial economics that uses econometric techniques to parameterize the relationships suggested. Financial mathematics is a field of applied mathematics, concerned with financial markets. The subject has a close relationship with the discipline of financial economics, which is concerned with much of the underlying theory that is involved in financial mathematics. Generally, mathematical finance will derive, and extend, the mathematical or numerical models suggested by financial economics. In terms of practice, mathematical finance also overlaps heavily with the field of computational finance. Arguably, these are largely synonymous, although the latter focuses on application, while the former focuses on modelling and derivation. The field is largely focused on the modelling of derivatives, although other important subfields include insurance mathematics and quantitative portfolio problems. Experimental finance aims to establish different market settings and environments to observe experimentally and provide a lens through which science can analyze agents' behavior and the resulting characteristics of trading flows, information diffusion and aggregation, price setting mechanisms, and returns processes. Researchers in experimental finance can study to what extent existing financial economics theory makes valid predictions and therefore prove them, and attempt to discover new principles on which such theory can be extended and be applied to future financial decisions. Research may proceed by conducting trading simulations or by establishing and studying the behavior, and the way that these people act or react, of people in artificial competitive market-like settings.

1) Summary generated by Singular Valued Decomposition

In economics, finance is a field that is concerned with the allocation (investment) of assets and liabilities over space and time, often under conditions of risk or uncertainty. Here, the twin assumptions of rationality and market efficiency lead to modern portfolio theory (the CAPM), and to the Black–Scholes theory for option valuation; it further studies phenomena and models where these assumptions do not hold, or are extended. "Financial economics", at least formally, also considers investment under "certainty" and hence also contributes to corporate finance theory. Generally, mathematical finance will derive, and extend, the mathematical or numerical models suggested by financial economics. Experimental finance aims to establish different market settings and environments to observe experimentally and provide a lens through which science can analyze agents' behavior and the resulting characteristics of trading flows, information diffusion and aggregation, price setting mechanisms, and returns processes. Researchers in experimental finance can study to what extent existing financial economics theory makes valid predictions and therefore prove them, and attempt to discover new principles on which such theory can be extended and be applied to future financial decisions.

2) Salience Score of each sentence

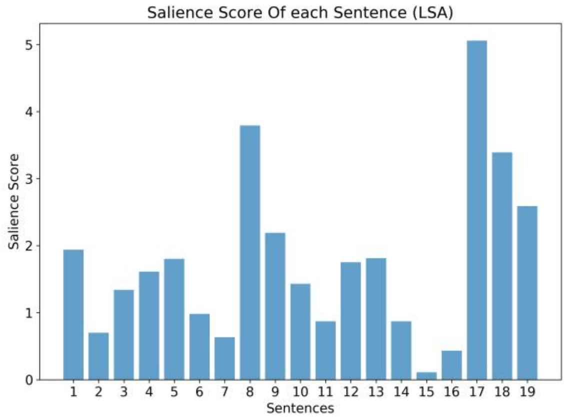

Figure 1: Salience score of each sentence computed by LSA

3) Summary generated by TextRank

Financial economics is the branch of economics studying the interrelation of financial variables, such as prices, interest rates and shares, as opposed to goods and services. It centres on managing risk in the context of the financial markets, and the resultant economic and financial models. "Financial economics", at least formally, also considers investment under "certainty" and hence also contributes to corporate finance theory. Financial mathematics is a field of applied mathematics, concerned with financial markets. The subject has a close relationship with the discipline of financial economics, which is concerned with much of the underlying theory that is involved in financial mathematics.

4) Text Rank Graph

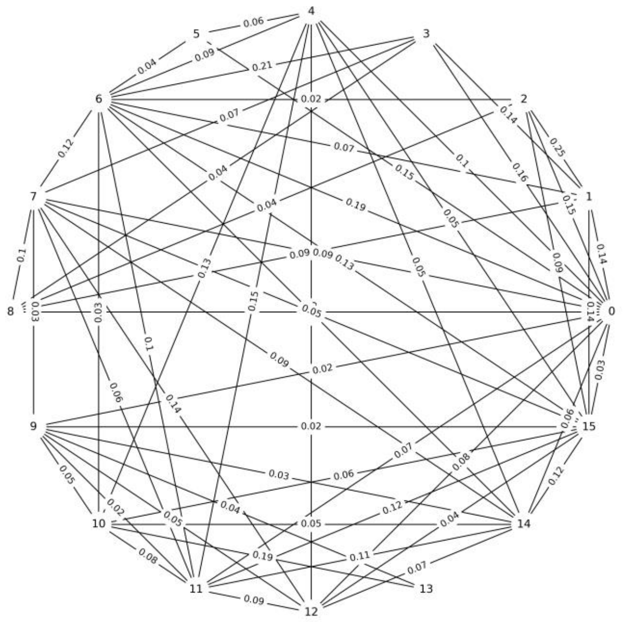

Figure 2: Text Rank Graph before score computation

5) TextRank Score of each sentence

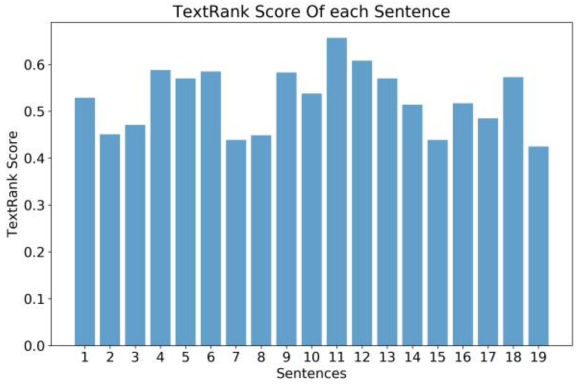

Figure 3: TextRank Score for each sentence in graph

6) Summary generated by Gensim

In economics, finance is a field that is concerned with the allocation (investment) of assets and liabilities over space and time, Participants in the market aim to price assets based on their risk level, fundamental value, and their expected rate of return. Financial economics is the branch of economics studying the interrelation of financial variables, such as prices, It centres on managing risk in the context of the financial markets, and the resultant economic and financial models. It essentially explores how rational investors would apply risk and return to the problem of an investment policy. "Financial economics", at least formally, also considers investment under "certainty" and hence also contributes to corporate finance theory. Financial mathematics is a field of applied mathematics, concerned with financial markets. Generally, mathematical finance will derive, and extend, the mathematical or numerical models suggested by financial economics. The field is largely focused on the modelling of derivatives, although other important subfields include insurance mathematics and quantitative portfolio problems. Experimental finance aims to establish different market settings and environments to observe experimentally and Researchers in experimental finance can study to what extent existing financial economics theory makes valid predictions and therefore prove them,

B. *Sample Text 2*

Cricket is a bat-and-ball game played between two teams of eleven players on a field at the centre of which is a 20-metre (22-yard) pitch with a wicket at each end, each comprising two bails balanced on three stumps. The batting side scores runs by striking the ball bowled at the wicket with the bat, while the bowling and fielding side tries to prevent this and dismiss each player (so they are "out"). Means of dismissal include being bowled, when the ball hits the stumps and dislodges the bails, and by the fielding side catching the ball after it is hit by the bat, but before it hits the ground. When ten players have been dismissed, the innings end and the teams swap roles. The game is adjudicated by two umpires, aided by a third umpire and match referee in international matches. They communicate with two off-field scorers who record the match's statistical information.There are various formats ranging from Twenty20, played over a few hours with each team batting for a single innings of 20 overs, to Test matches, played over five days with unlimited overs and the teams each batting for two innings of unlimited length. Traditionally cricketers play in all-white kit, but in limited overs cricket they wear club or team colours. In addition to the basic kit, some players wear protective gear to prevent injury caused by the ball, which is a hard, solid spheroid made of compressed leather with a slightly raised sewn seam enclosing a cork core which is layered with tightly wound string. Historically, cricket's origins are uncertain and the earliest definite reference is in south-east England in the middle of the 16th century. It spread globally with the expansion of the British Empire, leading to the first international matches in the second half of the 19th century. The game's governing body is the International Cricket Council (ICC), which has over 100 members, twelve of which are full members who play Test matches. The game's rules are held in a code called the Laws of Cricket which is owned and maintained by Marylebone Cricket Club (MCC) in London. The sport is followed primarily in the Indian subcontinent, Australasia, the United Kingdom, Ireland, southern Africa and the West Indies, its globalisation occurring during the expansion of the British Empire and remaining popular into the 21st century. Women's cricket, which is organised and played separately, has also achieved international standard. The most successful side playing international cricket is Australia, having won seven One Day International trophies, including five World Cups, more than any other country, and having been the top-rated Test side more than any other country.

1) Summary generated by Singular Valued Decomposition

Cricket is a bat-and-ball game played between two teams of eleven players on a field at the centre of which is a 20-metre (22-yard) pitch with a wicket at each end, each comprising two bails balanced on three stumps.There are various formats ranging from Twenty20, played over a few hours with each team batting for a single innings of 20 overs, to Test matches, played over five days with unlimited overs and the teams each batting for two innings of unlimited length.In addition to the basic kit, some players wear protective gear to prevent injury caused by the ball, which is a hard, solid spheroid made of compressed leather with a slightly raised sewn seam enclosing a cork core which is layered with tightly wound string.The sport is followed primarily in the Indian subcontinent, Australasia, the United Kingdom, Ireland, southern Africa and the West Indies, its globalisation occurring during the expansion of the British Empire and remaining popular into the 21st century.

2) Salience Score of each sentence

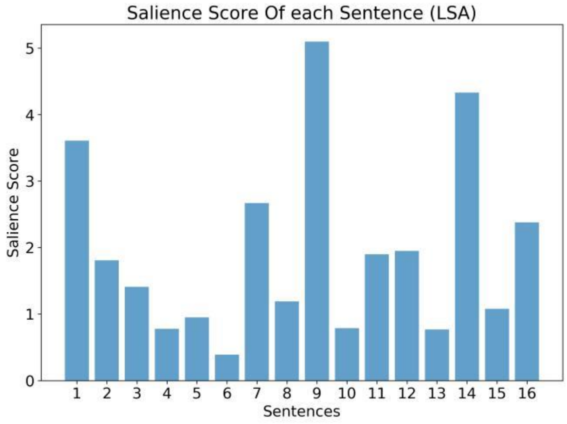

Figure 4: Salience score of each sentence computed by LSA

3) Summary generated by TextRank

Cricket is a bat-and-ball game played between two teams of eleven players on a field at the centre of which is a 20-metre (22-yard) pitch with a wicket at each end, each comprising two bails balanced on three stumps. The batting side scores runs by striking the ball bowled at the wicket with the bat, while the bowling and fielding side tries to prevent this and dismiss each player (so they are "out"). There are various formats ranging from Twenty20, played over a few hours with each team batting for a single innings of 20 overs, to Test matches, played over five days with unlimited overs and the teams each batting for two innings of unlimited length. The most successful side playing international cricket is Australia, having won seven One Day International trophies, including five World Cups, more than any other country, and having been the top-rated Test side more than any other country.

4) TextRank Graph

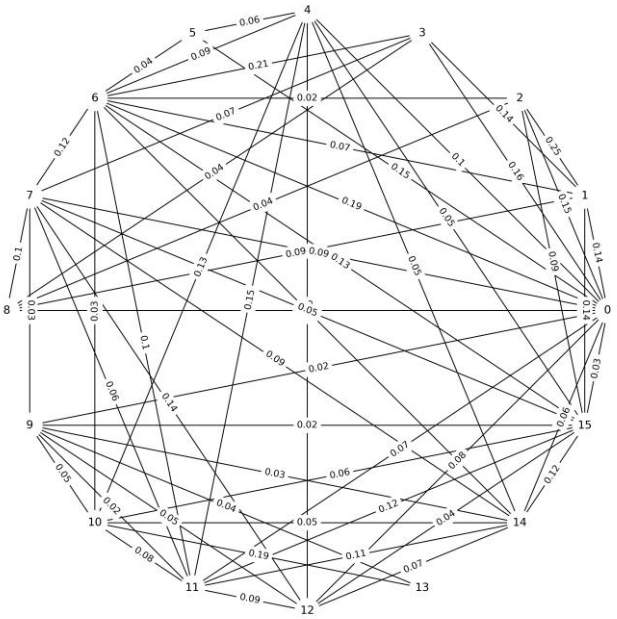

Figure 5: Text Rank Graph before score computation

5*)* TextRank Score of each sentence

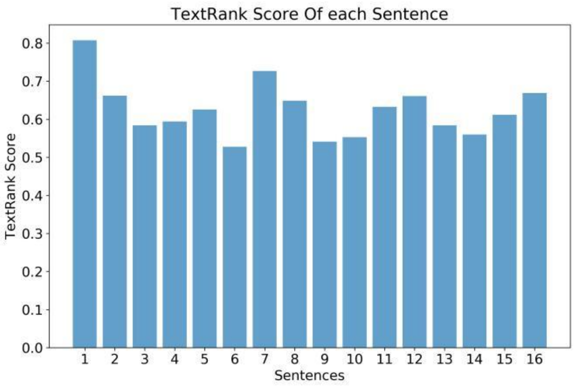

Figure 6: TextRank Score for each sentence in graph

6) Summary generated by Gensim

Cricket is a bat-and-ball game played between two teams of eleven players on a field at the centre of which is a 20-metre (22-yard) pitch with a wicket at each end, each comprising two bails balanced on three stumps. while the bowling and fielding side tries to prevent this and dismiss each player (so they are "out"). There are various formats ranging from Twenty20, played over a few hours with each team batting for a single innings of 20 overs, to Test matches, played over five days with unlimited overs and the teams each batting for two innings of unlimited length. Traditionally cricketers play in all-white kit, but in limited overs cricket they wear club or team colours. The game's governing body is the International Cricket Council (ICC), which has over 100 members, twelve of which are full members who play Test matches.

V. COMPARISON AND RESULTS

For comparing both methods we have used keywords. We extracted keywords of original text document and then analyzed whether these are captured by the summaries or not. The summary should be able to capture the main theme of the document and hence sentences containing keywords should be part of the summary.

We used collocations and weighted tag based keyword extraction methods for extracting top bigrams and trigrams present the document. We have also compared the summary generated by the two methods discussed in this paper with a text summarizer which is part of gensim library.

On comparing LSA and TextRank we found out that LSA generally runs faster than TextRank. A possible reason could be that in textrank we iterate over the graph many times until convergence. But LSA has underneath it the Singular Valued Decomposition which can be solved very fast. In case of how much keywords are captured by summary , LSA does a better job. In most of the text samples , the proportions of keywords was higher in LSA as compared to TextRank. In cases where a document contains many subtopics , the summaries generated by LSA was containing sentences from all these topics but this was not the case with textrank. This can be explained by the fact that in Singular Valued Decomposition , the singular matrix contains weights of all the major topics in the document. When this matrix is multiplied by right singular matrix we get scores for all the sentences. Now this multiplication favours sentences that contains words representing the major topics. Also one interesting observation is that there is less variation in the scores assigned by Text Rank but more variation in score assigned by LSA. This means LSA is better in seperating important sentences from less important sentences. When compared with the text summarizer present in gensim library , we found that there was more content overlap between gensim text summarizer and textrank as compared to LSA.

VI. CONCLUSION

Both graph based and matrix based methods have been compared on various parameters and results have been discussed. An effort has been been made to explain the results obtained. In future work , topic modelling methods will be incorporated to obtain summary that contains major topics present in the document. Such summaries are better concise representations of the entire document.

REFERENCES

[1] Karel Jezek and Josef Steinberger, "Automatic Text summarization", Vaclav Snasel (Ed.): Znalosti 2008, pp.1- 12, ISBN 978-80-227-2827-0, FIIT STU Brarislava, Ustav Informatiky a softveroveho inzinierstva, 2008. /1

[2] Weiguo Fan, Linda Wallace, Stephanie Rich, and Zhongju Zhang, “Tapping into the Power of Text Mining”, Journal of ACM, Blacksburg, 2005./4

[3] G Erkan and Dragomir R. Radev, “LexRank: Graph-based Centrality as Salience in Text Summarization”, Journal of Artificial Intelligence Research, Re-search, Vol. 22, pp. 457-479 2004./32

[4] Udo Hahn and Martin Romacker, “The SYNDIKATE text Knowledge base generator”, Proceedings of the first International conference on Human language technology research, Association for Computational Linguistics , ACM, Morristown, NJ, USA , 2001./33

[5] Farshad Kyoomarsi, Hamid Khosravi, Esfandiar Eslami and Pooya Khosravyan Dehkordy, “Optimizing Text Summarization Based on Fuzzy Logic”, In proceedings of Seventh IEEE/ACIS International

Conference on Computer and Information Science, IEEE, University of Shahid Bahonar Kerman, UK, 347-352, 2008./2

[6] Vishal Gupta, G.Sl Lehal, “A Survey of Text Mining Techniques and Applications”, Journal of Emerging Technologies in Web Intelligence, VOL. 1, NO. 1, 60-76, AUGUST 2009./31

[7] Jimmy Lin., “Summarization.”, Encyclopedia of Database Systems. Heidelberg, Germany: Springer-Verlag, 2009./46

[8] Jackie CK Cheung, “Comparing Abstractive and Extractive Summarization of Evaluative Text: Controversiality and Content Selection”, B. Sc. (Hons.) Thesis in the Department of Computer Science of the Faculty of Science, University of British Columbia, 2008./47

[9] G. Erkan and D. Radev. Lexrank: Graph-based centrality as salience in text summarization. Journal of Artificial Intelligence Research, 2004./25

[10] R. Mihalcea and P. Tarau. Textrank: Bringing order into texts. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, pages 404–411, 2004./61

[11] R. Mihalcea and P. Tarau. An algorithm for language independent single and multiple document summarization. In Proceedings of the International Joint Conference on Natural Language Processing, pages 19–24, 2005./62

[12] Y. Gong and X. Liu. Generic text summarization using relevance measure and latent semantic analysis. In Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 19–25, 2001. /33